**1. Word count (excluding headers)**

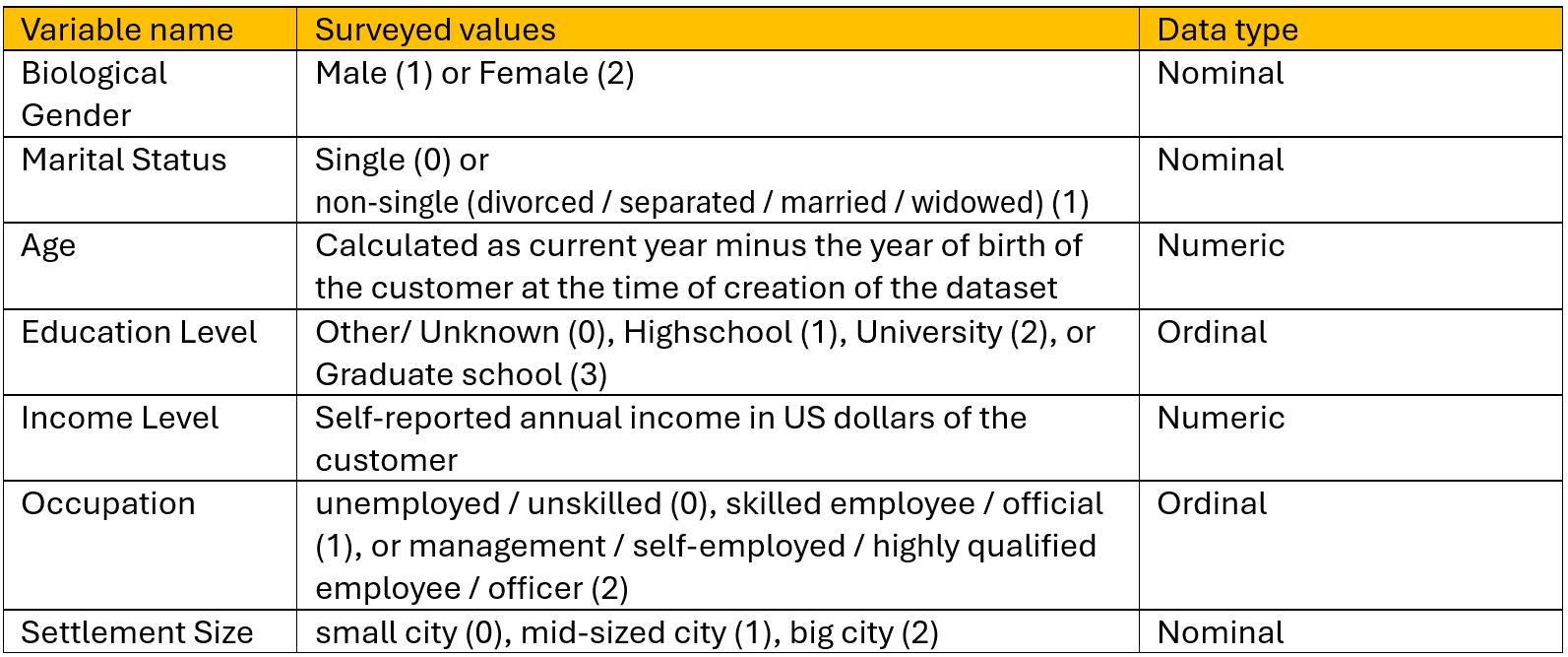
*-* 1062 words

**2. Introduction**

The objective of this report is to segment traveling customers using a quantitative approach based on a demographic dataset with 2000 observations across six variables: Biological Gender, Marital Status, Age, Education Level, Income Level, Occupation, and Settlement Size. Unsupervised machine learning techniques, specifically clustering, are used due to the absence of predefined target labels. The methodology includes Exploratory Data Analysis (EDA), standardization of numeric variables, and determining the optimal number of clusters using the Elbow method and Silhouette charts. Next, 2 clustering techniques-K-means++ (choosing centroids first and locating neighboring data) and Agglomerative clustering (treating each observation as a cluster and merging them)- are applied. Finally, the report suggests tailored marketing strategies for each segment.

**3. Exploratory Data Analysis**

Table 1: Variable description



The general statistics of the dataset can be seen from the table below:

A screenshot of a computer screen

Description automatically generated

Preliminary statistics show a dominant number of respondents are male, non-single, 41 years old, having a high school graduate, earning an average of $137,516 USD annually, unemployed, and living in small cities. This indicates that major customers of this brand are from a lower to medium social background, implying that the data presenting is from a commodity product.

The age distribution of the customer is right-skewed (most frequent age of 33, mean age of) 41, suggesting that the brand attracts a relatively young demographic. Together with the commodity insight, this may mean that this product is more tailored to youngsters (maybe due to health issues or preferences)

A graph of a distribution of age

Description automatically generated

Similarly, the income distribution is right-skewed, suggesting that the brand's products likely appeal to medium-income individuals, potentially positioning them as essential or commodity-based items that provide value across income levels. High-income customers, though less frequent, may also find value in the brand’s offerings, suggesting a broad appeal with potential for upselling premium options.A graph of a distribution of income

Description automatically generated

Similarly, the income distribution is right-skewed, suggesting that most customers have medium-to-low incomes. This conclusion aligns with the notion that the brand’s products are likely targeted at a broad audience, valuing affordability and convenience. Moreover, the negative relationship between income and demand reinforces the hypothesis that the service may be considered inferior, as its consumption seems to decrease as income levels rise.

**4. Customer Segmentation**

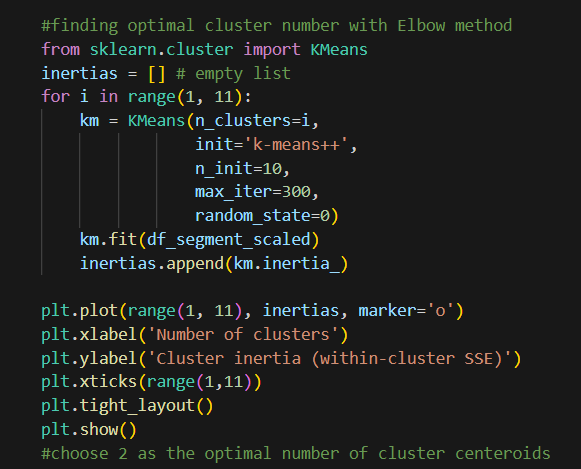
The first step for customer segmentation is standardization for numeric variables- Age and Income:

A computer screen shot of text

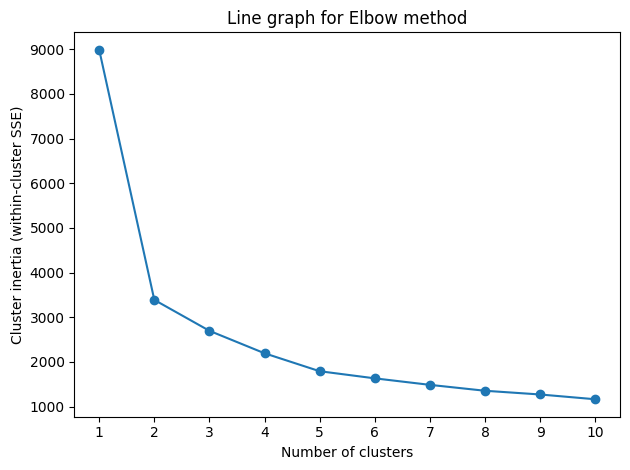
Description automatically generated

Secondly, we will need to determine the optimal number of clusters using Elbow method and Silhouette Plots.

Firstly, Elbow methods determine the number of clusters (k) based on inertia attribute (sum of squared distances of samples to their closest cluster center). Specifically, k is chosen where inertia stops decreasing rapidly. The code is:



The output is:



Based on the model output, the report writer chooses 2 as the optimal number of clusters due to its highest slope, indicating the highest change in cluster inertia.

Silhouette charts are another way to determine optimal cluster number. The idea is to go through some numbers of clusters and choose the one with the highest Silhouette Coefficient. Under this scenario, we will create 3 Silhouette graphs with the number of clusters of 2,3 and 4. The code is:

A screen shot of a computer program

Description automatically generated

The outputs are:

A graph of a graph

Description automatically generated with medium confidence

A graph of a graph

Description automatically generated with medium confidence

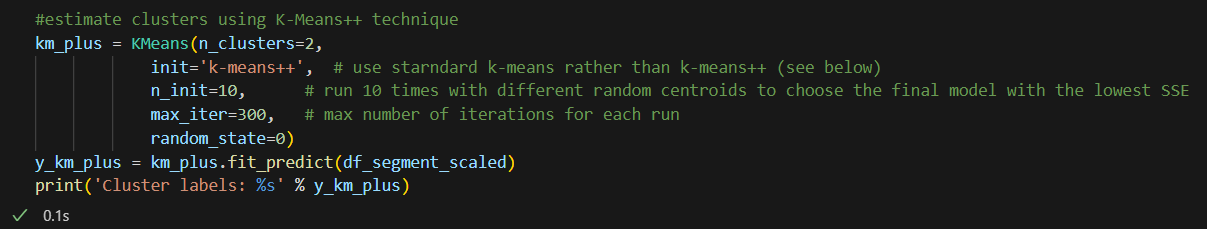
A graph of different colors

Description automatically generated

Based on the charts, 2 conclusions are made. Firstly, all 3 clusters achieve Silhouette coefficients not close to 0, indicating that the dataset has been appropriately clustered. Secondly, coefficients are decreasing when k increases, meaning that the Silhouette coefficient is the largest at k=2. Thus, we can conclude that k=2 is the optimal number of clusters, matching with the decision of Elbow Method.

After choosing segmentation numbers, we will start to cluster. The process starts with choosing the initial cluster centroids, which can be done in 2 ways- K-means++ (top-down approach) and Agglomerative Clustering (bottom-up approach). Firstly, K-Means++ chooses initial centroids sequentially such that they have high probability of being far away from each other (not belong in the same cluster).

The code is:



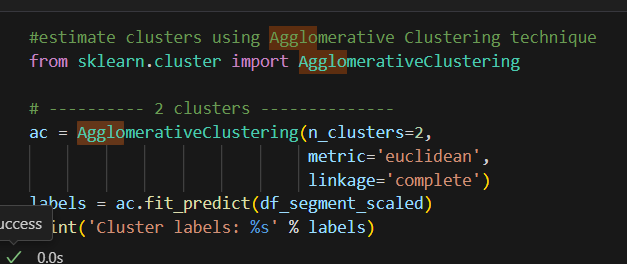
The output is:

A black and white text

Description automatically generated

On the other hand, Agglomerative clustering follows the steps of treating each data point as an individual cluster and iteratively merges the two closest clusters, updating the distance matrix, and repeating until all points are combined into one cluster.

The code is:



The output is:

A black background with white text

Description automatically generated

To get the cluster centroids, we will compute the representative data of each variable for each cluster 0 and 1. Specifically, Income and Age will be treated with average value while the other variables are treated with the most frequent data (mode). The tables below summarize the cluster centers and number of customers in each cluster:

For K-means++:

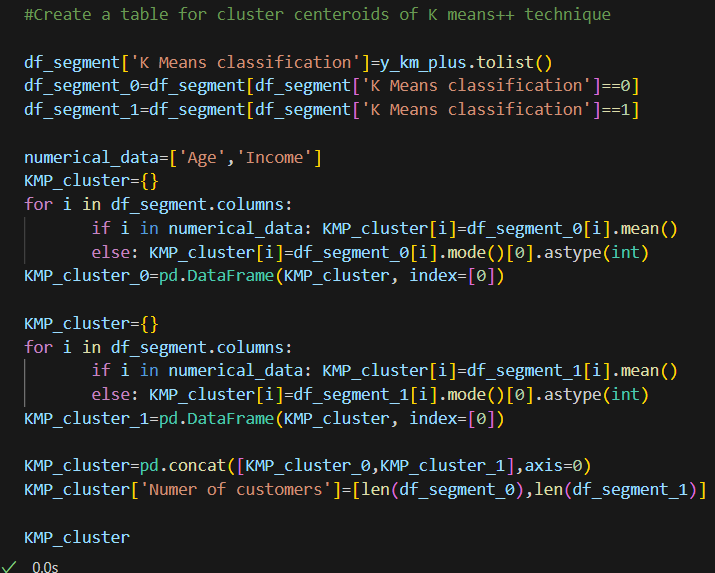
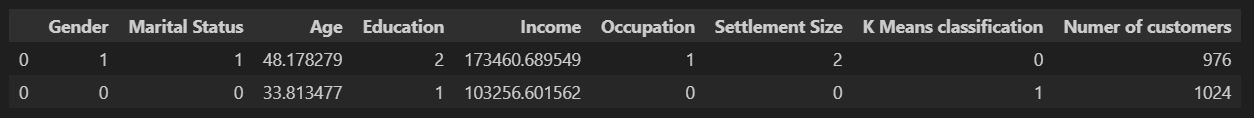


Table 2: Cluster centroids for K-means++

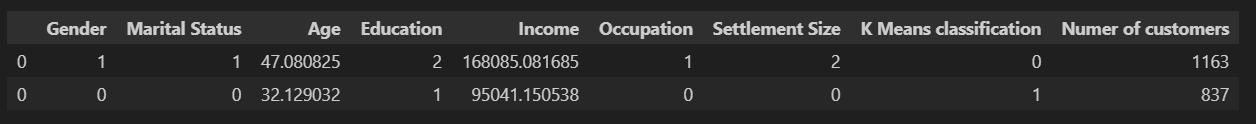


For Agglomerative clustering:

A computer screen shot of text

Description automatically generated

Table 3: Cluster centroids for Agglomerative clustering



Based on 2 table outputs, the profile of each cluster for each technique is:

Table 4: Cluster centroids for K-means++

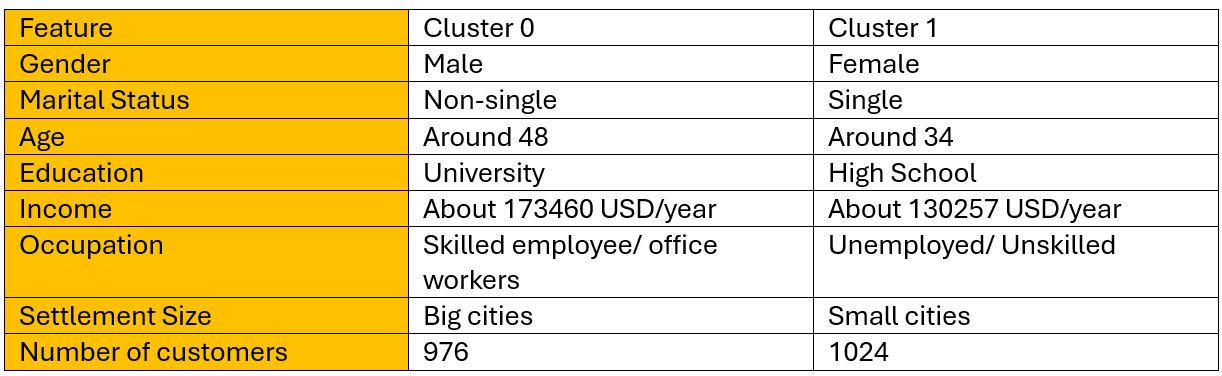


Table 5: Cluster centroids for Agglomerative clustering

A screenshot of a computer

Description automatically generated

We can see that there are a lot of overlaps between clustering using the 2 techniques, especially among non-numerical variables. Specifically, there are a total of 1813 overlap results between 2 techniques, indicating there are minor differences among the clustering results.

The overlap summary:

A black screen with white text

Description automatically generated

**6. Recommendations**

For segment 0, a premium positioning strategy is the most effective. Marketing should emphasize luxury feelings and family comfortability when traveling. Channels like LinkedIn and professional financial blogs are well-suited to this segment, given their high engagement with professional content. Besides, Exclusive offers, loyalty programs, and partnerships with luxury brands can reinforce the service's premium image. Moreover, Messaging should focus on responsibility, success, and wealth accumulation, aligning with their goals of family security and stability.

In contrast, segment 1 requires a more cost-centered marketing approach. The strategy should emphasize affordability and accessibility to travelling. Platforms such as Instagram and TikTok are ideal for engaging this group through relatable content and influencer partnerships. Promotions like bulk pricing, flexible payment plans and referral programs are highly recommended. Furthermore, messaging should center on empowerment, independence, and financial control, reflecting this segment’s need for budget-friendly solutions. Collaborations with local organizations or community events will further build trust and connection, especially in smaller urban settings, making the product more approachable.

**7. Conclusion**

In conclusion, the segmentation analysis using K-means++ and Agglomerative Clustering has highlighted two distinct customer groups with similar market demand yet differing in needs and preferences. By targeting the older, high-income segment through knowledge-driven channels and addressing their higher-level needs, while engaging the younger, average-income segment through social media and cost-focused messaging, businesses can effectively reach both groups. This tailored approach ensures that marketing strategies align with each segment's unique characteristics, maximizing engagement and driving long-term success.